

TripActions

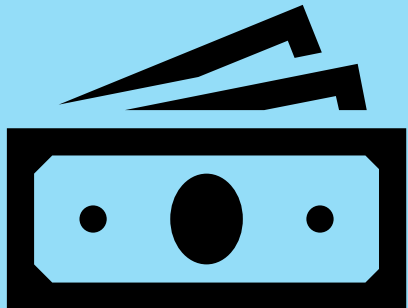
Predicting Flight Delays

Group 2B

"The Amazing Butterflies"

DDT & A.I. II

16.04.20

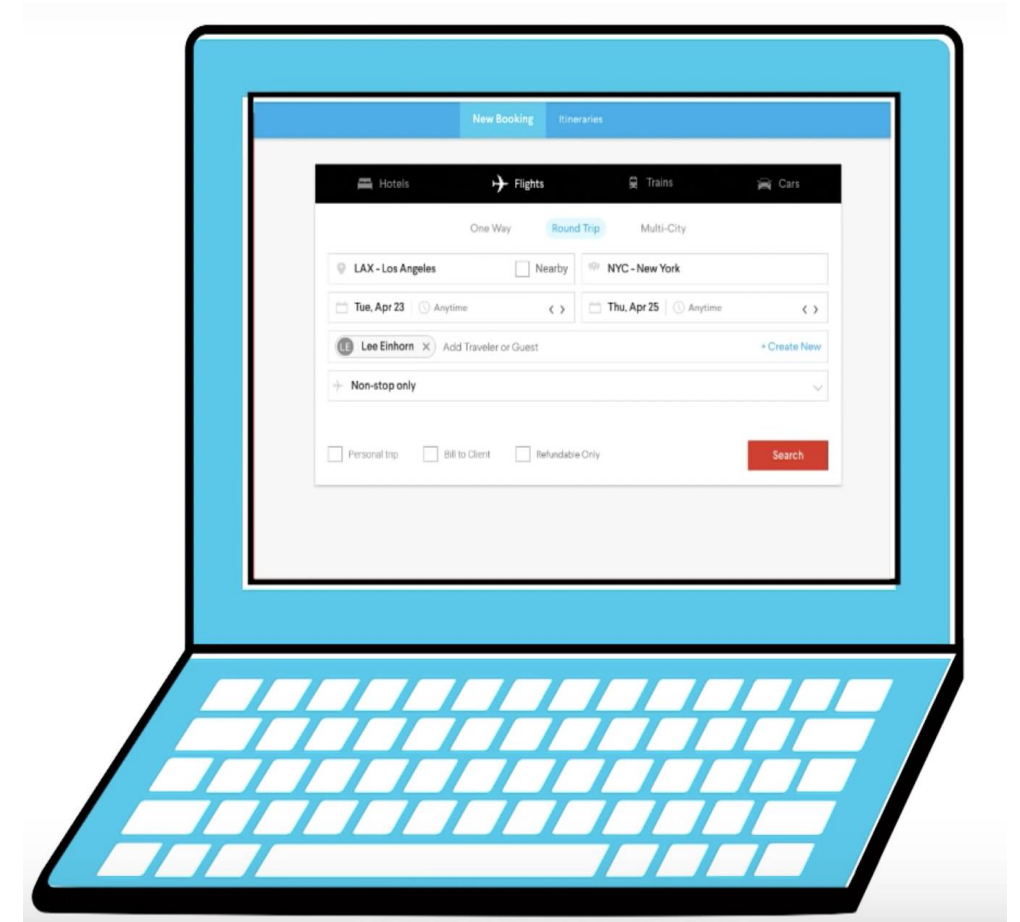


TripActions

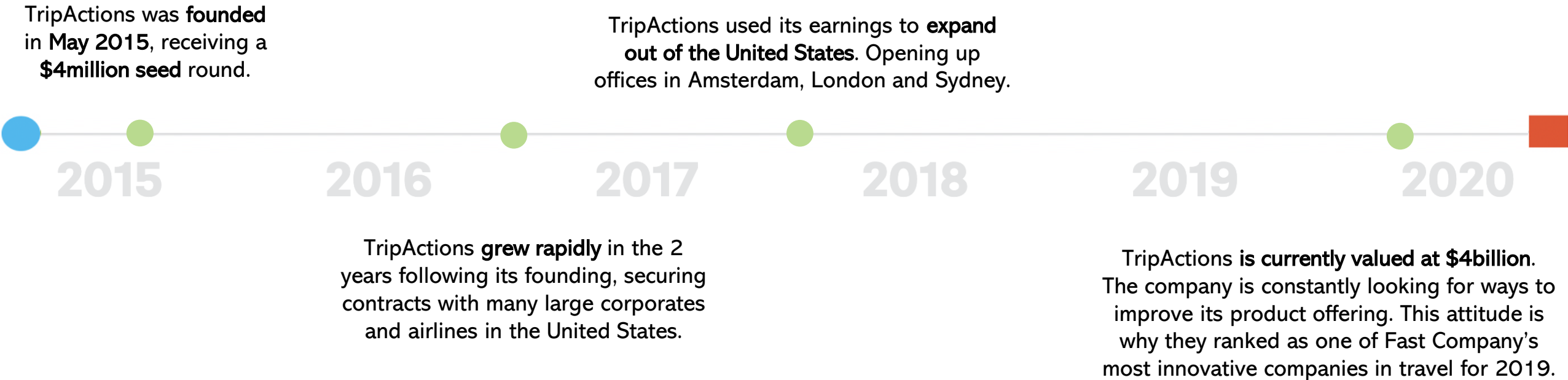
A platform that takes care of business and your travelers.

TripActions ensures business travel continuity and the safety of your traveling employees.

Using the right data and developing useful insights, TripActions provides a platform that enables companies to build, manage and scale a world-class corporate travel management program with ease.

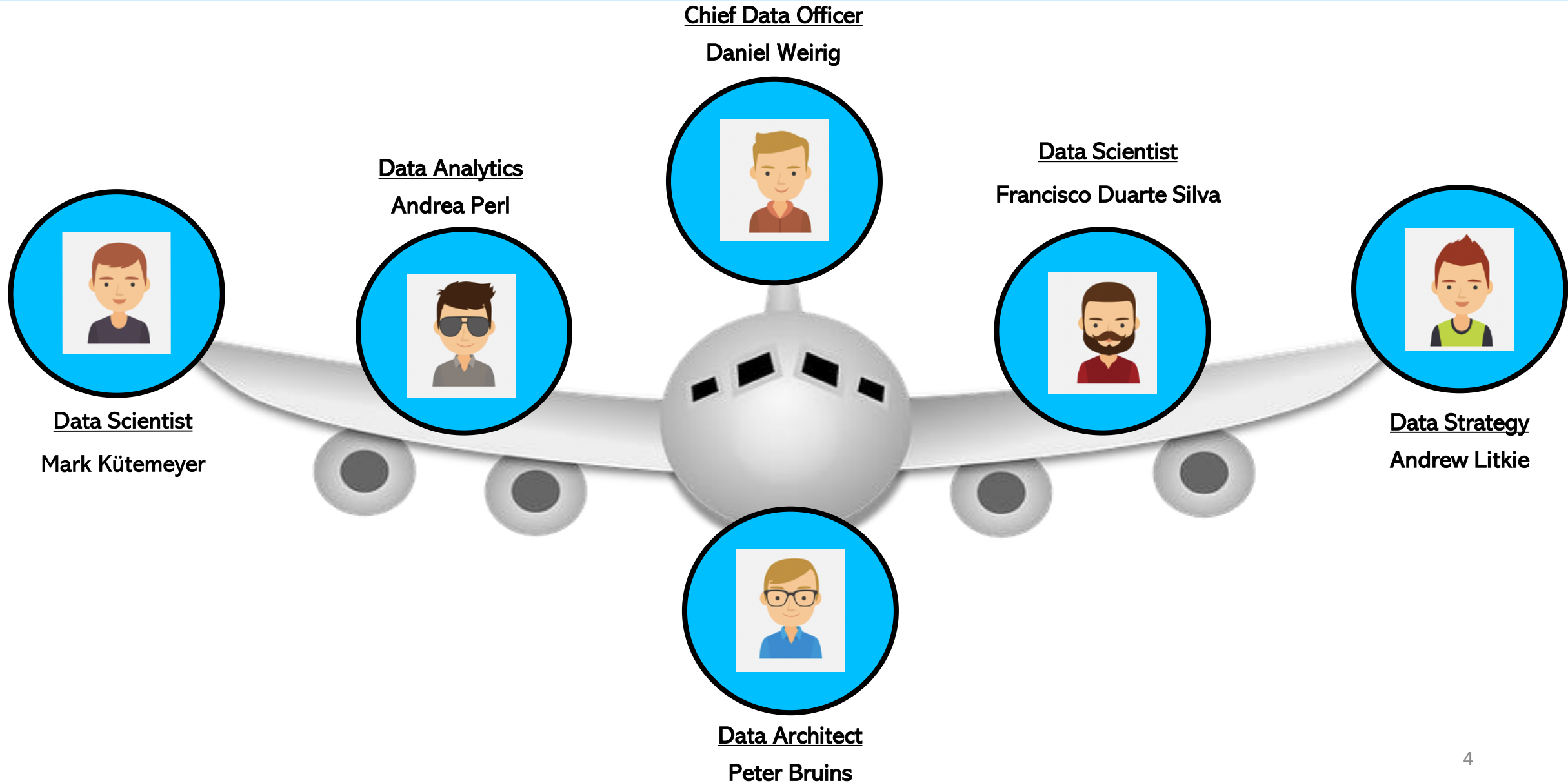


TripActions' Story

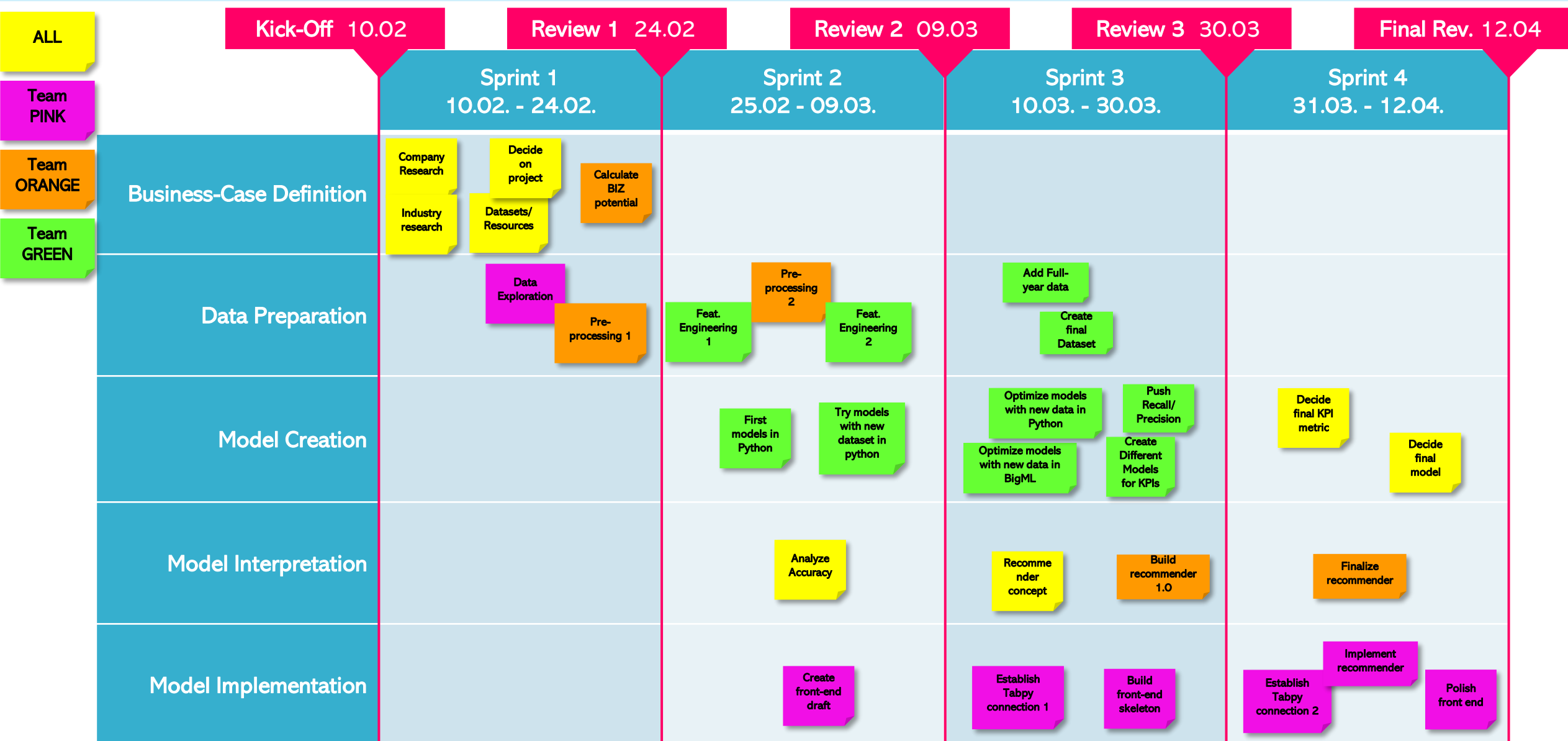


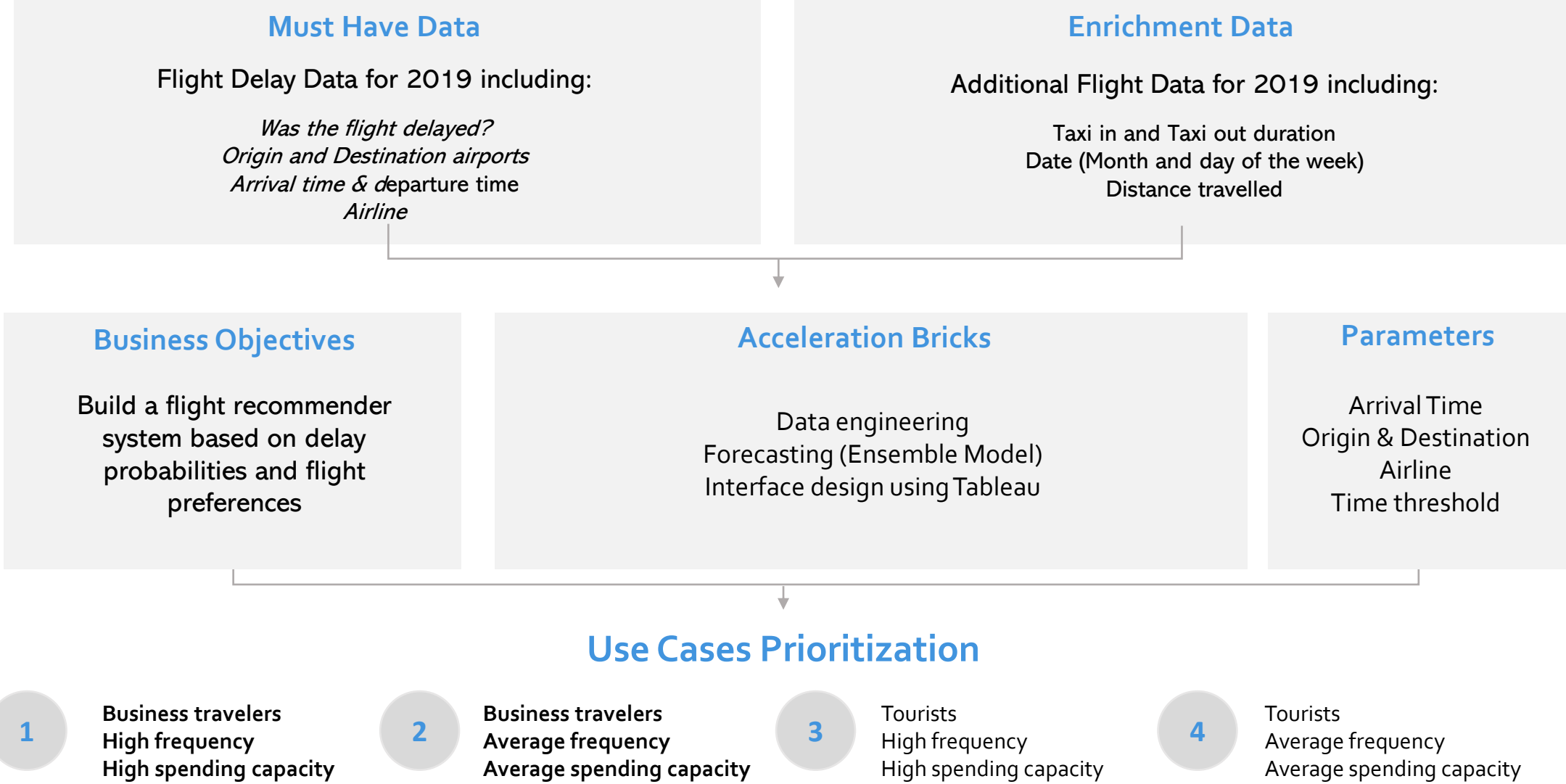
Ariel Cohen, the CEO of TripActions, has recently announced that he is **employing the help of a data driven transformation company, known as The A Team**, in order to help TripActions stay ahead of competition.

The A Team



A Project in 4 Sprints





Description of Opportunity

Travel management platforms currently have no indication for customers on arrival delays. As a fast growing startup, TripActions can use a flight delay predictor to be the go-to travel management platform for business travel and differentiate itself from.

Industry

Online travel management

Value / ROI Calculation

Revenue per corp. customer	\$ 200K
Revenue estimated at \$20 million (2019)	
Company has around 1000 customers (2019)	
Increase in customers	5%
Value generated	\$ 1M
Investment	\$ 450K
Team of six developers	
Avg. salary per developer in California is 75K	
ROI (first year)	220%

Model

We are aiming to build a classification model, predicting delays using 10+ variables regarding flight and airline data

Context

Travel management companies will need to differentiate in light of corona pandemic

Implementation

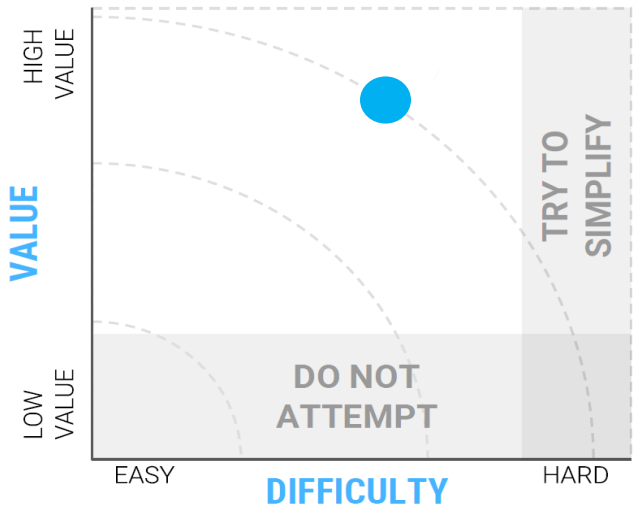
We are aiming to build a recommender system showing flights with low chances of delay according to several criteria:

1. Desired arrival time and flexibility (hour threshold)
2. Date
3. Origin and destination airport
4. Desired Airline

System will be built using Tableau

Opportunity Size

Global business travel expected to rise to \$1.6 trillion by 2020



Source: Owler, TravelPerk



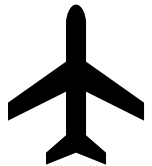
Merge monthly 2019 flights data from the US Bureau of Transportation Statistics



Explore the data and drop NA's



Change departure time to departure hour & date to day of week to extract value

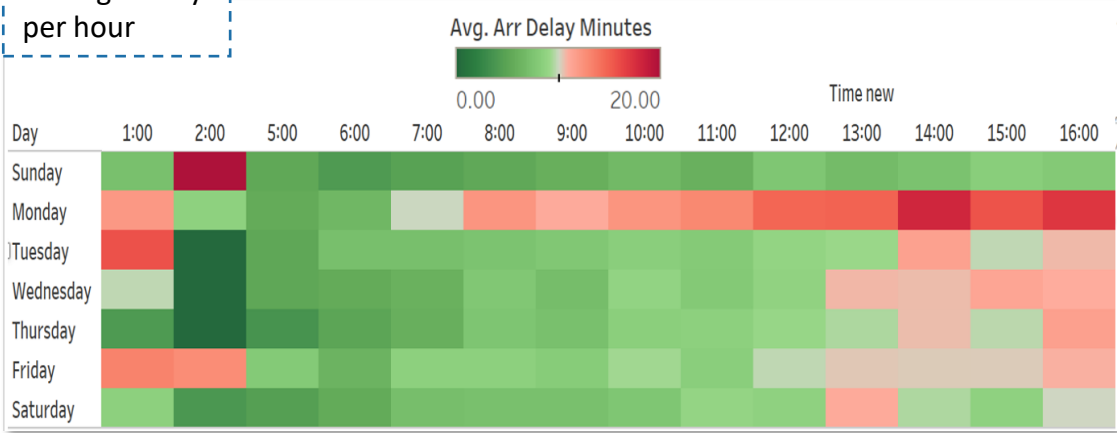


Consider top 10 busiest airports for business case

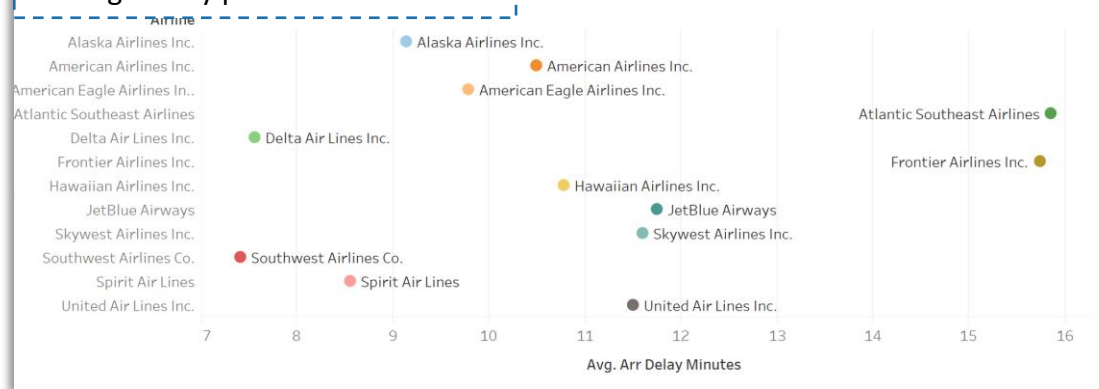


Define the binary target: Delays
Delayed if ≥ 15 minutes behind schedule

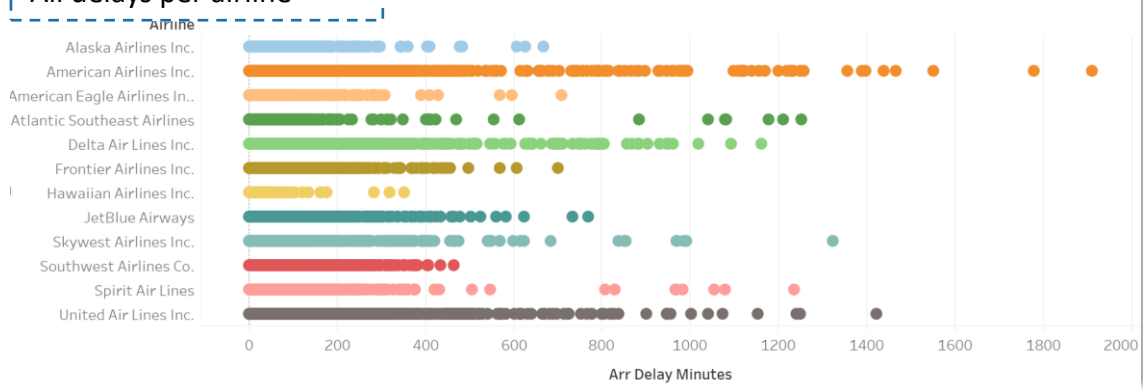
Average delay
per hour



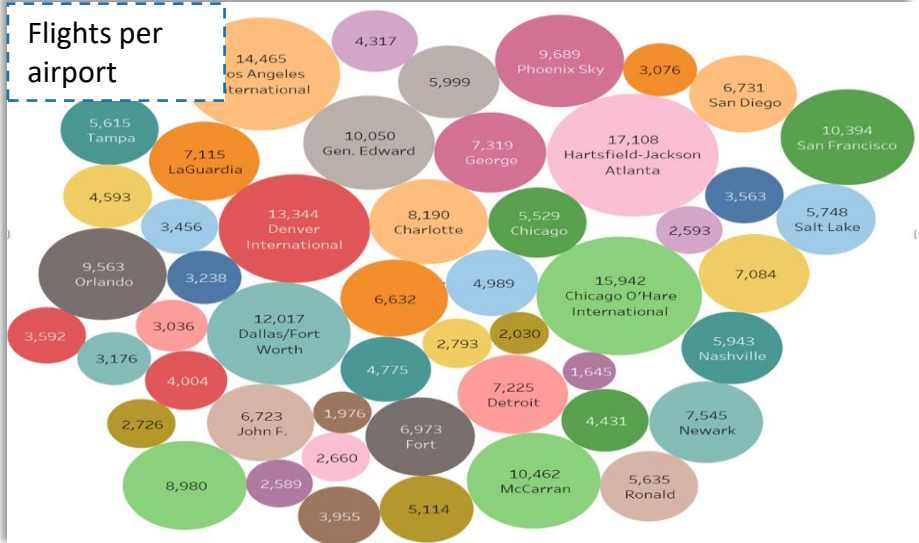
Average delay per airline



All delays per airline



Flights per
airport



Data set contains over 150 features

Drop 120 features which are not relevant for project

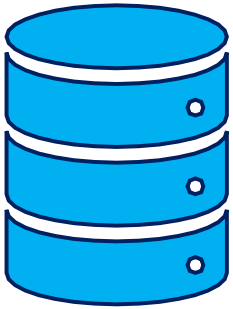
Drop 20 features based on tableau analyses and model tests

Create new features based on insights

Use anomaly detection to delete 500 biggest anomalies from train data

Let's Build the model!





1 target

4 numerical features

6 categorical features

475,706 observations

MONTH	ABC	475,706	0	0	
DAY_OF_WEEK	ABC	475,706	0	0	
OP_CARRIER	ABC	475,706	0	0	
ORIGIN	ABC	475,706	0	0	
DEST	ABC	475,706	0	0	
ARR_DEL15	ABC	475,706	0	0	
CRS_DEP_TIMEHour	123	475,706	0	0	
Speed	ABC	475,706	0	0	
avg_tax_out	123	475,706	0	0	
avg_tax_in	123	475,706	0	0	
airline_delay	123	475,706	0	0	

Building the Model

We attempted several models with Python and BigML

Trees

Random
Forest



Boosted
Trees

XgBoost



Deep
Learning

Ann



Bootstrap
Forest



Lightgbm



Catboost



Run on Python



Run on BigML

We ran several models through Jupyter Notebooks

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import GridSearchCV
from sklearn.metrics import make_scorer

cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=0) sampler = SMOTomek(sampling_strategy=1, random_state=42, n_jobs=-2)

rnf = RandomForestClassifier(random_state=random_state)
scoring = {'f1_score': make_scorer(f1_score),
          'f1_score': make_scorer(f1_score)}

params = {'max_depth': [15, 20, 25, 30],
          'min_samples_split': [5, 7, 10],
          'min_samples_leaf': [2, 4, 6],
          'n_estimators': [150, 400, 450]}

grid_clf = GridSearchCV(estimator=rnf, param_grid=params, cv=cv,
                        scoring=scoring, n_jobs=-2, verbose=10,
                        shuffle=True, random_state=0)
grid_clf.fit(X_train, y_train)

best_params = grid_clf.best_params_
best_model = grid_clf.best_estimator_

model = RandomForestClassifier(**best_params)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

# Evaluating the model
from sklearn.metrics import classification_report
print(classification_report(y_test, y_pred))
```

We also used the BigML platform

	random2 Training (80%) top 746 anomalies dataset 33 total models (deepnet: 2, ensemble: 31), metric=precision for ...		ARR_DEL15	6d	1h 9min
	random2 Training (80%) top 746 anomalies dataset 33 total models (deepnet: 2, ensemble: 31), metric=recall for 1.0, ...		ARR_DEL15	6d 10h	1h 1min
	random2 Training (80%) 33 total models (deepnet: 2, ensemble: 31), metric=recall for 1.0, ...		ARR_DEL15	6d 11h	1h 3min
	random2 Training (80%) 33 total models (deepnet: 2, ensemble: 31), metric=precision for ...		ARR_DEL15	6d 21h	1h 23min
	random2 Training (80%) AA 57 total models (deepnet: 2, ensemble: 55), metric=precision for ...		ARR_DEL15	6d 21h	1h 35min
	random2 Training (80%) 8 total models (deepnet: 1, ensemble: 7), metric=precision for 0.0, ...		ARR_DEL15	6d 21h	12min

Model Evaluation

Which metrics should we opt for?

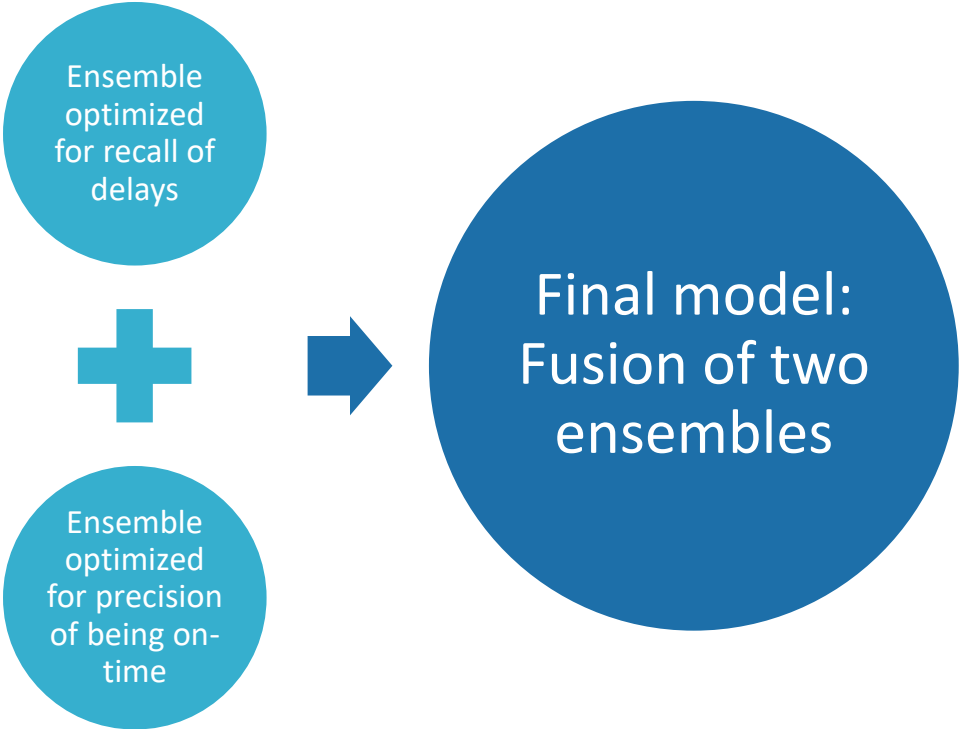
- 1 Recall: Minimize False negatives
True class would be delay
- 2 Precision: Minimize False positives
True class would be on-time

Metric Results

	Recall	Precision	AUC
True Class Delays	64.4%	85.2%	68.3%
True Class Non-delays	68.9%	85.4%	68.2%



Final Model



ACTUAL VS. PREDICTED

	1.0	0.0		
1.0	14,818	7,089	21,907	67.64%
0.0	29,540	43,695	73,235	59.66%
PREDICTED	44,358	50,784	95,142	63.65% AVG. RECALL
PRECISION	33.41%	86.04%	59.72% AVG. PRECISION	61.50% ACCURACY

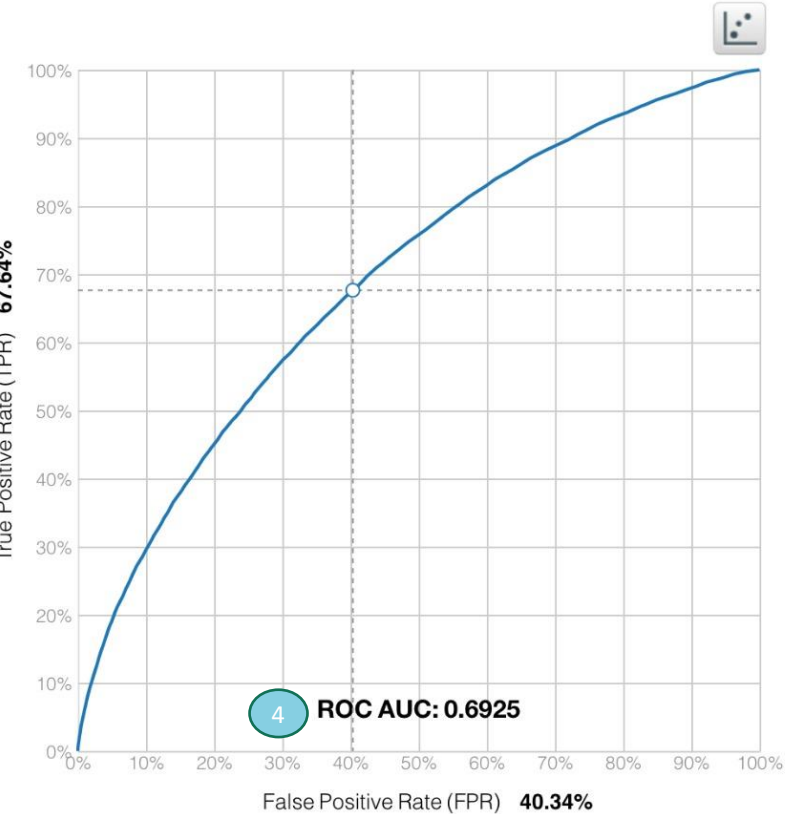
TP

FN

FP

TN

F	Phi
0.45	0.23
0.70	0.23
0.58 AVG. F	0.23 AVG. Phi

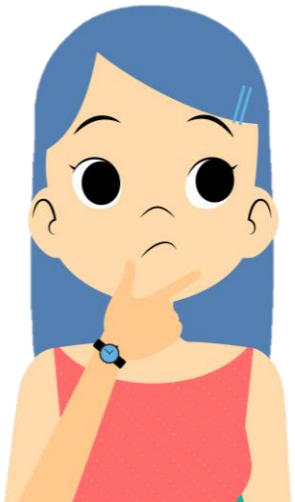


61.5% Accuracy	0.4472 F-measure	
33.4% Precision	67.6% Recall	0.2304 Phi coefficient
40.3% FPR	46.6% % positive instances	145.1% Lift
27.6% K-S statistic	0.2313 Kendall's Tau	0.2833 Spearman's Rho

- 1 False Negatives (7,089)
- 2 Recall Metric (67.6%)
- 3 Precision Metric (86 %)
- 4 AUC Metric (0.70)

... from a recommendation system I want to:

- 1 Be shown the flights with the lowest chance of being delayed...
- 2 ... not arriving later than a certain time ...
- 3 ... but also not arriving too early ...
- 4 ... and preferably from my favorite airline



Taking into account the customer's specifications, we have built a recommender system:



Using python for the hidden workings of the system

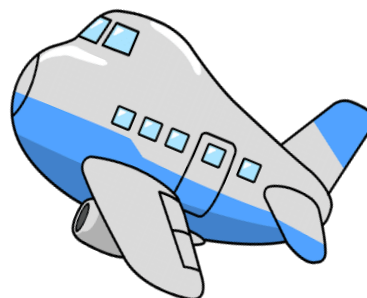


Using Tableau to build a user-friendly interface



The goal is to smoothly and successfully integrate the delay predictor into TripActions business model.

The premium and luxury columns are empty. Why?
Not available for domestic flights!



Therefore these two columns can be replaced to
show the airline and the probability of delay.

TripActions New Booking [Need Help?](#) Rewards \$0 Hi Sean

Departure Date: 2020/11/20 Arrival Time: 13:30 Flexibility (hours): 2 Origin: Chicago Destination: San Francisco

Departure Flights

Q Flight # Any stops Time Airlines Price Policy Recommended

Fare Categories	Standard	Enhanced	Airline	Probability of Delay
Median prices by column	\$690	\$903		
10:31 am → 13:16 pm Nonstop	Main Cabin (L) Round Trip from \$170	Comfort+ (S) Round Trip from \$334	American Airlines	31.24%
10:08 am → 12:53 pm Nonstop	Main Cabin (L) Round Trip from \$152	Comfort+ (S) Round Trip from \$347	United Airlines	33.32%
09:55 am → 12:52 pm Nonstop	Main Cabin (L) Round Trip from \$142	Comfort+ (S) Round Trip from \$301	American Airlines	34.85%
08:52 am → 12:00 pm Nonstop	Main Cabin (L) Round Trip from \$133	Comfort+ (S) Round Trip from \$272	Delta Airlines	42.35%

Putting the Recommender to use



A business traveller wants to book a flight on the 20th of November 2020.

Wanting to arrive at 13:30 pm latest and her schedule allows for flexibility of arriving 2 hours before this time.



And travelling from Chicago to San Francisco.



TripActions TripActions New Booking Trips Need Help? Rewards \$0 Hi Sean

Departure Date: 2020/10/01 Arrival Time: - Flexibility (hours): - Origin: - Destination: -

Departure Flights

Q Flight # Any stops Time Airlines Price Policy Recommended

Fare Categories	Standard	Enhanced	Airline	Probability of Delay
Median prices by column ①	\$690	\$903		
Nonstop → 	Main Cabin (L) Round Trip from	Comfort+ (S) Round Trip from		
Nonstop → 	Main Cabin (L) Round Trip from	Comfort+ (S) Round Trip from		
Nonstop → 	Main Cabin (L) Round Trip from	Comfort+ (S) Round Trip from		
Nonstop → 	Main Cabin (L) Round Trip from	Comfort+ (S) Round Trip from		



TripActions

Thank you.
Do you have any questions?





TripActions











Back-Up.





[Github with the whole project](#)

The ML Canvas is now complete

The machine learning canvas				
Designed for: Tripaction		Designed by: The A team		Date: 13-04 Iteration: 4
Decisions  How are predictions used to make decisions that provide the proposed value to the end-user?	ML task  Input, output to predict, type of problem.	Value Propositions  What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?	Data Sources  Which raw data sources can we use (internal and external)?	Collecting Data  How do we get new data to learn from (inputs and outputs)?
Precision and recall vs accuracy Minimum precision for no delay prediction of 77% required	Binary classification based on multiple numerical and categorical features Unbalanced dataset True/False ratio 2:7	Predict chance of flights being delayed or not, to make sure business traveler are on time as much as possible	U.S. Bureau of Transportation Statistics (part of United States Department of Transportation)	Visualize features in tableau to gain wide overview Input features with slight changes to understand relation with delay
Making Predictions  When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?	Offline Evaluation  Methods and metrics to evaluate the system before deployment.	Model outperforms random decision of always no delay prediction, which is right 77% of the time. Added value is difference between model and 77%.	Features  Input representations extracted from raw data sources.	Building Models  When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?
Model receives prediction request from recommender system which is imbedded into tableau to create a live feedback experience to userbase	<ul style="list-style-type: none"> Cost of False Negative Cost of False Positive Minimum benchmark accuracy 		<ul style="list-style-type: none"> Airport Airline Time <ul style="list-style-type: none"> DotW Hour Month Taxi in/out time Etc. 	Start with data from one month and optimize model for the data Input data of whole year and make slight adjustments
		Live Evaluation and Monitoring  Methods and metrics to evaluate the system after deployment, and to quantify value creation.		
		Update database with new monthly data and check validity		

Key takeaways

- Understand the implications of Precision/ Recall on the business case and its implication to the tradeoffs being made in the model
 - How "expensive" a false negative is
 - How "expensive" a false positive is
 - What is minimum accuracy required for the product
- Know how to cope with unbalanced data

Putting the Recommender to use e.g. #2



A business traveller wants to book a flight on the 2nd of November 2020.

She wants to arrive at 20:30 pm latest and her schedule allows for flexibility of arriving 2 hours before this time, but no earlier.



She is travelling from from Dallas to Denver. The reason for this trip is for an important meeting with a new client.

TripActions

New Booking

Trips

Need Help?

Rewards \$0

Hi Sean

Departure Date

2020/11/02

Arrival Time

20:30

Flexibility (hours)

2

Origin

Dallas

Destination

Denver

Departure Flights

Q Flight #

Any stops

Time

Airlines

Price

Policy

Recommended

Fare Categories	Standard	Enhanced	Airline	Probability of Delay
Median prices by column	\$690	\$903		
18:54 pm Nonstop 20:00 pm	Main Cabin (L) Round Trip from \$72	Comfort+ (S) Round Trip from \$143	American Airlines	27.12%
18:56 pm Nonstop 20:12 pm	Main Cabin (L) Round Trip from \$82	Comfort+ (S) Round Trip from \$156	American Airlines	27.12%
17:31 pm Nonstop 18:37 pm	Main Cabin (L) Round Trip from \$63	Comfort+ (S) Round Trip from \$110	United Airlines	30.13%
17:55 pm Nonstop 19:01 pm	Main Cabin (L) Round Trip from \$52	Comfort+ (S) Round Trip from \$110	Spirit Airlines	50.18%

How the integration was so smooth

1

To integrate the recommender system into TripActions front end, we decided to leverage the power of TabPy and Tableau.

4

To establish the Taby and Tableau connection you need to open up the TabPy and get the port number, as seen below:

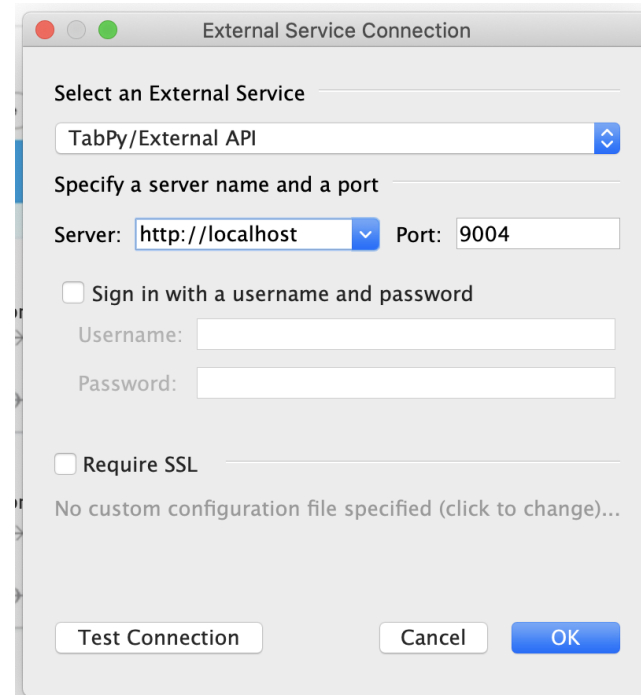
```
Last login: Mon Apr 13 07:56:06 on console
tabpy
The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Andrews-MacBook-Pro-2:~ andrewlitkie$ tabpy
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter port set to "9004" from
default value
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter server_version set to "1
.0.0" from default value
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter evaluate_timeout set to
"30" from default value
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter upload_dir set to "/User
s/andrewlitkie/opt/anaconda3/lib/python3.7/site-packages/tabpy/tmp/query_objects
" from default value
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter transfer_protocol set to
"http" from default value
2020-04-13,12:15:54 [DEBUG] (app.py:app:212): Parameter certificate_file is not
set
2020-04-13,12:15:54 [DEBUG] (app.py:app:212): Parameter key_file is not set
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter state_file_path set to "
/Users/andrewlitkie/opt/anaconda3/lib/python3.7/site-packages/tabpy/tabpy_server
" from default value
2020-04-13,12:15:54 [DEBUG] (app.py:app:206): Parameter static_path set to "/Use
```

2

The following needed to be installed:
- TabPy (pip install tabpy)
- Virtual Environment (pip install virtualenv)

5

Then on Tableau you need to manage the connection show that Tableau is listening to the port that was just opened up:



3

Create a virtual environment for TabPy to be able to connect to Tableau:
- virtualenv my-tabpy-env

6

Then you need to open up the recommender system on Jupyter notebook through the virtual environment that was created:

```
andrewlitkie — -bash — 80x24
Last login: Mon Apr 13 12:15:51 on ttys000

The default interactive shell is now zsh.
To update your account to use zsh, please run `chsh -s /bin/zsh`.
For more details, please visit https://support.apple.com/kb/HT208050.
(base) Andrews-MacBook-Pro-2:~ andrewlitkie$ source my-tabpy-env/bin/activate
(my-tabpy-env) (base) Andrews-MacBook-Pro-2:~ andrewlitkie$ ipython notebook
```

7

The recommender system needs to be deployed into the TabPy environment. This is done on the right, which uses the recommender system to create an endpoint in the localhost called FlightPredictor.

```
# Connect to TabPy server using the client library
connection = tabpy_client.Client('http://localhost:9004/')

# Create recommender system:

def recommend(arrival_time, origin, destination, airline, threshold, rec_data = data, model = model):

# Publish the function to TabPy server so it can be used from Tableau
# Using the name Flight Predictor it can be called on Tableau
connection.deploy('FlightPredictor',
                  recommend,
                  'Returns probability of delay of flights in the US', override = True)
```

8

By searching the following URL: <http://localhost:9004/endpoints>, you can observe all the deployed functions. Where you can see that the the FlightPredictor function has been successfully deployed.

```
localhost:9004/endpoints

{"GermanCreditCheck": {"description": "Classifies bad or good according to the model trained by relevant dataset", "type": "model", "version": 6, "dependencies": [], "target": null, "creation_time": 1585140925, "last_modified_time": 1586270551, "schema": null, "docstring": "-- no docstring found in query function --"}, "DiagnosticsDemo": {"description": "Returns diagnosis suggestion based on ensemble model trained using Wisconsin Breast Cancer dataset", "type": "model", "version": 12, "dependencies": [], "target": null, "creation_time": 1585143498, "last_modified_time": 1586078702, "schema": null, "docstring": "-- no docstring found in query function --"}, "DiagnosticsDemo1": {"description": "Returns diagnosis suggestion based on ensemble model trained using Wisconsin Breast Cancer dataset", "type": "model", "version": 1, "dependencies": [], "target": null, "creation_time": 1585315750, "last_modified_time": 1585315750, "schema": null, "docstring": "-- no docstring found in query function --"}, "FlightPredictor": {"description": "Returns probability of delay of flights in the US", "type": "model", "version": 13, "dependencies": [], "target": null, "creation_time": 1585841412, "last_modified_time": 1586270353, "schema": null, "docstring": "-- no docstring found in query function --"}}
```

9

The last step is to call the function in Tableau. To do this you need to create a calculated field to query the endpoint that was just created.

Flight Predictor

Results are computed along Table (across).

```
SCRIPT_STR("return tabpy.query('FlightPredictor',_arg1,_arg2,_arg3,_arg4,_arg5,_arg6)['response']",
[date], [Time], [origin], [destination], [airline], [threshold])
```

The calculation is valid.

Default Table Calculation

Apply

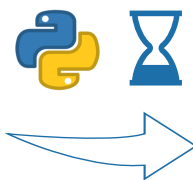
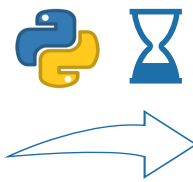
OK

How do we deliver?



I have a dinner with a client at 9pm in Denver, but I can only check in at my hotel from 6:30pm

... also, most of my miles are with American Airlines!!!



Origin: DFW
Destination: DEN

	Date	Departure Time	Arrival Time	Airline	Probability of Delay
6957952	28/12/2020	18:56	20:12	AA	31.07%
6906869	28/12/2020	17:25	18:32	AA	35.93%
6996924	28/12/2020	17:52	18:58	AA	35.93%
6996931	28/12/2020	17:21	18:33	AA	35.93%
7255138	28/12/2020	17:55	19:01	NK	53.41%
7251505	28/12/2020	18:15	19:25	NK	60.63%
7255143	28/12/2020	18:00	19:07	NK	60.63%

Origin: DFW
Destination: DEN
Airline: AA

	Date	Departure Time	Arrival Time	Probability of Delay
6957952	28/12/2020	18:56	20:12	31.07%
6906869	28/12/2020	17:25	18:32	35.93%
6996924	28/12/2020	17:52	18:58	35.93%
6996931	28/12/2020	17:21	18:33	35.93%